

**<<Hadoop-MapReduce>>**

**工作量估计与统计分析**

项目成员

SY1506107 陈昆度

SY1506112 马宇晴

SY1506105 杨云龙

SY1506423 李嘉艺

北京航空航天大学

2016-6-12

目录

[1． 任务概要 3](#_Toc452740824)

[1.1 实验内容概述 3](#_Toc452740825)

[1.2 计划和分工 6](#_Toc452740826)

[2． 具体任务分配 7](#_Toc452740827)

[2.1软件开发计划书 7](#_Toc452740828)

[2.2软件需求说明书 7](#_Toc452740829)

[2.3软件测试需求说明书 8](#_Toc452740830)

[2.4软件制作 8](#_Toc452740831)

[2.5配置管理 8](#_Toc452740832)

# 任务概要

## 实验内容分配概述

在整个软件工程实施的过程当中，我们所有的实验内容和任务大概如下图所示。首先是软件开发计划书。

然后是软件需求计划书。

然后是软件制作。

然后是测试需求说明。

然后是测试过程。

然后是工程管理。

## 计划和分工

首先是人员大致分工情况，在组内马宇晴主要负责文案类型的工作，李嘉艺负责环境配置和代码审核，陈昆度负责管理类的工作，杨云龙负责代码的工作。但是在实际过程中考虑到个人能力的问题工作出现了一些交叉，但大致上的安排没有大变化。具体的安排会在第二章中详细说明。

整个计划的安排大致可以见Mpp。总的来说任务的完成情况还可以，大部分任务都按照时间及时完成。

# 具体任务分配

## 2.1软件开发计划书

字数和内容统计：

|  |  |  |
| --- | --- | --- |
| 人员 | 内容 | 字数 |
| 马宇晴 | 引言、项目概述，背景部分的撰写 | 624字 |
| 陈昆度，杨云龙 | 实施计划草案，初步制定每周方案 | 人均504字 |
| 李嘉艺 | 其他部分的撰写 | 696字 |

参与版本变更情况：

|  |  |  |  |
| --- | --- | --- | --- |
| 人员 | 参与版本 | 内容 | 字数 |
| 马宇晴 | V1.0 | 负责统计审核意见 | 94字 |
| 杨云龙 | V1.0，1.01 | 修改文档形式部分 | 167字 |
| 陈昆度 | V1.0，1.01，1.02，1.03 | 修改背景，细化整理文档 | 146字 |
| 李嘉艺 | V1.0 | 负责会议记录 | 43字 |

可以从表中看出，在项目计划书的订正过程中添加的字数是很少的，这里需要提到一点是，由于我们做的第一个最原始版本完全不符合要求，所以我们直接略过没有作为第一版。

## 2.2软件需求说明书

字数和内容统计：

|  |  |  |
| --- | --- | --- |
| 人员 | 内容 | 字数 |
| 马宇晴 | 撰写和项目背景等有关内容 | 340字 |
| 杨云龙 | 撰写数据词典和缩略词，以及软件描述 | 475字，2张表 |
| 陈昆度 | 撰写RUCM和环境说明 | 356字，26张图 |
| 李嘉艺 | 撰写RUCM和功能需求 | 634字，8张图 |

参与版本变更情况：

|  |  |  |  |
| --- | --- | --- | --- |
| 人员 | 参与版本 | 内容 | 字数 |
| 马宇晴 | V1.0 | 负责统计审核意见，同时对其他组提出汇总意见 | 422字 |
| 杨云龙 | V1.0，1.01，1.02，1.03 | 修改一些细节问题 | 320字 |
| 陈昆度 | V1.0，1.01，1.02，1.03，1.04，1.05 | 针对评审以及互评意见，总体修改文档，细化文档。 | 更新15张RUCM图，修改字数752 |
| 李嘉艺 | V1.0，V1.01 | 负责RUCM图部分的修改 | 修改6张RUCM |

## 2.3软件测试需求说明书

字数和内容统计：

|  |  |  |
| --- | --- | --- |
| 人员 | 内容 | 字数 |
| 陈昆度 | 主要负责RUCM部分的书写 | 绘制9张图 |
| 李嘉艺 | 其他部分的书写 | 911字 |

参与版本变更情况：

|  |  |  |  |
| --- | --- | --- | --- |
| 人员 | 参与版本 | 内容 | 字数 |
| 马宇晴 | V1.0 | 负责统计审核意见，同时对其他组提出汇总意见 | 1678字 |
| 杨云龙 | V1.0，1.01 | 增加了用例的工时 | 9个表格 |
| 陈昆度 | V1.0，1.01，1.02 | 细化文档，修改一些有问题的部分 | 更新9张RUCM图，增加缩略词和软件测试环境 |
| 李嘉艺 | V1.0，V1.01 | 负责设计问题清单，同时按照意见进行修改 | 新增1张表 |

## 2.4软件制作

|  |  |  |
| --- | --- | --- |
| 人员 | 内容 | 代码行数 |
| 马宇晴 | 负责开发环境的搭建 | 配置信息42行 |
| 杨云龙 | 负责核心代码的开发，如核心函数的书写 | 493行 |
| 陈昆度 | 负责源代码修改，以及爬虫和预处理的撰写 | 130行 |
| 李嘉艺 | 负责代码汇总和审核 | 73行 |

## 2.5配置管理

|  |  |  |
| --- | --- | --- |
| 人员 | 内容 | 具体数据 |
| 马宇晴 | 负责会议记录，课堂意见，以及管理表格设计，还有其他各类文档制作 | 制作表格9张，字数2005 |
| 杨云龙 | 负责核心代码的开发，如核心函数的书写 | 制作PPT一次 |
| 陈昆度 | 负责mpp的管理和Github的整理。同时填写组员工作相关信息 | 制作PPT6次。 |
| 李嘉艺 | 负责修改和核实错误代码 | 制作PPt一次，搭建相关环境 |

## 2.6小结

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 工作评估 | 人员 | 工作量 | | | 工作评价 | 贡献率 | 说明 |
| 文字 | 代码 | 图表（包括PPT） |
| 陈昆度 | 1758+3436 | 130 | 56 | 工作认真负责，组织能力强，任务分配得当，在团队中能够起到带头作用 | 25% | 工作评估为小组互评 |
| 马宇晴 | 5163+2052 | 123 | 9 | 注重细节，文笔流畅 | 25% |
| 杨云龙 | 1466+2010 | 493 | 12 | 编程能力强，态度认真 | 25% |
| 李嘉艺 | 2284+2111 | 239 | 16 | 灵活机动，项目中的各个方面均有涉猎，发挥着不可替代的作用 | 25% |
| 新增/调整工作 | 工作名称 | 工作类型 | | | 工作量 | 变更原因 |  |
| 无 | 无 | | | 无 |  |
|
|
| 修改情况 | 修改次数 | 修改量 | | | 修改工作量 | 修改原因 |  |
| 软件开发计划书3次 | 313字 | | | 8工时 | 格式不符合要求，预期不符合实际或描述不清 |
| 软件需求计划书5次 | 1494字，21张图表 | | | 24工时 | 复评审后接受老师和同学的意见修改，或因后期需求无法测试而修改 |
| 软件测试需求计划3次 | 1678字，19张图表 | | | 24工时 | 复评审后接受老师和同学的意见修改，或对照需求计划书更改 |
| 工作产出 | 项目开发计划书，软件需求计划书，软件测试需求计划书，软件测试报告，总结报告，问题清单，管理报告，会议记录 | | | | | | 如问题报告，测试分析报告等 |

# 有效方法分析

## 3.1 实验一：软件需求分析

最大的困难主要还是不清楚自己能做什么，要做什么，因为挑选出题目的时候其实对Hadoop系统并不熟悉，所以可能导致许多描述上的模糊，同时一开始也难以统一意见形成一个具体的目标。

目的：解决由于对系统不熟悉，难以确定软件实际需求的问题。

依据：为了了解我们能够做到的上限，就是制定一个既能够满足我们能力范围之内，同时又能实际做出效果的目标。我们主要查找内容包括Hadoop的技术手册和各类型的技术博客。但是最重要的是内部人员的多次讨论和沟通，确定应该在软工作业上花费多少时间。

人员：李嘉艺之前的软件工程经验比较丰富，是主要的问询对象，其他人在解决问题中的职责在于查早资料和熟悉Hadoop。

处理步骤：由于最大的分歧还是集中在是否要真的改进Hadoop的性能上，（因为数据集寻找和处理是必须要的），资料说明了网上许多公司以及个人都有修改Hadoop代码并且使之快速的先例，但是缺乏非常具体的手把手教学。因此确定是否修改性能是在没有现成教程或者示例的前提下，由组员商讨得到的意见。

结果：我们的需求主要内容是修改源码和数据处理。

## 3.2 实验二：软件需求评审

评审阶段主要分为对他人的评审以及评审之后对自身的修改。关于这个阶段本质上没有什么困难，但是存在一个版本更迭之后不知道和上一个版本之间有什么区别的问题。

目的：解决版本之间变更统计的问题。

依据：关于这点主要是需要依赖一些工具帮助我们进行这个工作，有许多文档编辑器自带这种功能，比如word。

人员：陈昆度负责研究如何利用word当中的功能构建出一个自己的版本管理方式。

处理步骤：主要利用了审阅这个功能，对每个修改的地方进行划线和批注，在进行版本更迭的时候，按照其他小组提出的意见接受上一版的意见，在写下一版本的审阅意见，文件中看来一目了然。

结果：使用了批注审阅的方式管理版本的区别。

## 3.3 实验三：软件产品改进与展示

其实这个阶段就是软件产品制作的阶段，这个阶段遇到的问题大多数比较隐私性，就是说这些问题都是属于我们小组独有的问题。而且多数都是纯粹的技术问题。有一个软工方面的问题是，怎么分工？

目的：解决人员分工问题。

依据：分工主要需要考虑到两个点，人员的工作量和能力。由于Hadoop系统我们在开始项目之前都没有接触过，所以我们对于Hadoop的理解程度也完全不同。最后在进行项目任务分配的时候，起到决定性作用的还是我们对系统理解的深浅决定了我们的分工。

人员：陈昆度对Hadoop较为了解，阅读了源码，看了较多文献。杨云龙主要了解Java语法，对Hadoop了解程度一般。李嘉艺认真了解Hadoop的用法，清楚调试流程，马宇晴擅长Hadoop的安装和配置，能够进行故障排除。

处理步骤：在各自阅读了较多的关于Hadoop的文献之后，我们讨论了每个人擅长的工作，并且据此分配了工作。在实际过程中显然会存在某个人能力不足以解决某个问题的情况，这种时候我们会集中起来进行会议讨论。

结果：根据每个人能力的不同，在进行Hadoop系统制作的时候我们的分工并没有按照工作量的方式进行区分，而是按照每个人擅长的点去处理。

## 3.4 实验四：软件测试

关于软件测试有一个很大的疑问，就是应该测试什么东西。或者说，文档之类的都很容易像模像样的就写出来了，但是其实在软件制作阶段的时候，如果有问题早就看出来并且修改了，实际过程中真的还会遇到测试问题吗？

目的：如何才能检测出可能存在的BUG？

依据：因为测试的时候存在许多方法，比如黑盒测试之类的，但是由于我们的项目本质上说非常小，所以测试时最需要注意的就是各类的用例测试点一定要找准，需要思考极端情况下可能发生的问题。

人员：关于测试我们其实都没有什么经验。

处理步骤：主要考虑的是应该尽量去测试之前在功能实现阶段时不会去想的一些极端状况，比如说价格区间的设定，设定为3000-5000可能不会出现问题，但是如果5000-10000呢，位数改变的时候字符串处理会不会出现偏差？按照这个思路我们进行了一系列的测试，最后的确找到了一些存在的BUG。

结果：测试的过程中发现了一些特殊情况，包括偶然数据错误，某个函数编译失败等问题。

## 3.5 实验五：软件测试评审

测试评审的过程中主要内容在于修改自己的测试内容，而测试这个事情非常绝对，也很封闭，测试结果，成功与否，大多数信息都在我们自己组内，相互能够提出的意见只能基于文本上。我们主要被询问的最多的一个问题是：表述不明确。

目的：为何会有人认为我们文本表述不明确？怎么解决

依据：所谓的表述不明确，举个例子，我们说，爬取数据需要验证，验证结果是数据正确。模糊的点在于正确是什么标准，到底怎么样算正确。

人员：负责撰写说明书RUCM的是陈昆度，用例测试RUCM是杨云龙，李嘉艺和马宇晴。

处理步骤：可能在某些点上我们自己认为一些描述的词汇非常多余，但是对于评审者来说可能十分重要，我们尽量改善了描述的方法，比如成功配置了环境，变成了使用-java和-maven命令来看监视器中是否会出现版本信息作为应答。

结果：修正了很大一部分的用例描述。

## 3.6 实验六：软件进度计划与控制

进度计划控制最大的问题集中在mpp上，关于mpp主要有几个问题，一个是没有及时记录的信息如何处理，第二个则是mpp如何使用。

目的：怎样正确的运用mpp帮助进度计划控制。

依据：mpp确实是能够很好的帮助计划控制，最好的方式应该是每周五之后立刻使用mpp制定下周的计划分配人员，然后组长使用mpp进行管理，但是目前我们没有这么做。

人员：mpp主要由陈昆度负责。

处理步骤：关于mpp的使用，从不熟悉到理解了几个比较重要的功能，比如说基线控制可以很好帮助版本管理，工时和工期可以通过日历进行自动调整。同时今天知道了不能给父类任务设置资源

结果：明白了如何用mpp控制进度。

## 3.7 实验七：工作量估计和统计分析

工作量估计工作非常困难，主要是因为复杂度和工作量之间的关系不好处理，还有统计分析也是一个难点。

目的：如何从工作量统计数据中看出什么结论。

依据：和常规的数据分析不一样，如果按照人事的角度，应该最后会分析出来一个谁去谁留或者谁工作能力强，谁领导能力弱，项目组织的好不好之类的东西，但是我不确定我们的实验应该使用这样的切入点。

人员：马宇晴，李嘉艺负责制作表格，陈昆度负责总结。

处理步骤：最后我们利用数据总结出来的是我们的分工情况和贡献率，以及大家最擅长的工作是什么。

结果：主要从人员的角度进行了项目的分析。

## 3.8 实验八：配置管理

关于配置管理其实主要使用的工具应该是github以及之前提到的通过审批的方式进行版本跟踪。

目的：如何使用github进行版本管理。

依据：一些非常正规的github项目，比如说mysql，docker等大的厂家纷纷使用github进行版本管理。

人员：李嘉艺主要负责上传和整理github。

处理步骤：一开始我们对github的文件结构等都非常模糊，不知道该如何使用，甚至上传下载都不会，但是随后在一些交流中，我们发现使用离线文件常常会出现接收不到或者接收不及时，或者服务器删除文件的情况，此事我们在github上就可以很容易的找到最新的文件，只需要一个约定好的规则，就非常易于使用。

结果：使用github进行文件的管理，下载和交流。

# 4.分析与总结

首先可以看到的是，在我们组分配任务的过程中，较为严格的按照了分工的进行了作业，从工作类型上面来看，主要的工作类型分为文案类，图表类，管理类和程序类。其中陈昆度和李嘉艺的主要任务是管理和图表工作，马宇晴的工作主要是文案类，杨云龙的工作主要是代码类。

从工作的复杂度上看，代码和图表工作复杂度较高，管理和文案类工作比较简单。总的来说，简单的工作特征在于比较繁重，而复杂的工作则比较要求能力。组长陈昆度有软件开发经验，同时负责给团队出谋划策。所以主要工作都在于管理方面。

李嘉艺之前有过比较多的软件工程经验，所以可以帮助RUCM图表的制作以及管理工作，同时能够胜任一部分代码职责。

马宇晴工作速度比较快，同时比较细心，所以在团队中其主要负责配置管理和数据统计，同时参与文案的写作。

杨云龙比较精通各类编程语言，因此主要参与了核心代码的写作。

总的来看，我们组内的分工比较得当，发挥了每个人作用的同时工作量分配的也很均匀。

由于我们的工作也主要在于文档写作和软件开发，因为实际难度并不算非常的高，主要困难还是集中于能够实现功能同时改良hadoop的代码。所以在整个过程中执行时主要困惑都在于能够正常的完成我们的目标，而不是文案性质的东西是否足够规范。所以屡次导致我们的各类文档修改次数过多，然而软件开发的结果却比较成功，从最开始预定的目标开始到最后，没有做出大幅度的目标修改，最后却顺利的完成了目标，同时实现了排序算法和Hadoop的性能改进。所以我们最后的优缺点都非常明显，缺点在与我们的软件工程有关内容文档不够正规，优点则在于我们成功的利用了软件工程的方法统筹规划，顺利的完成了内容。

最后，有关于更多的分析总结内容，可以参见我们的PPT。